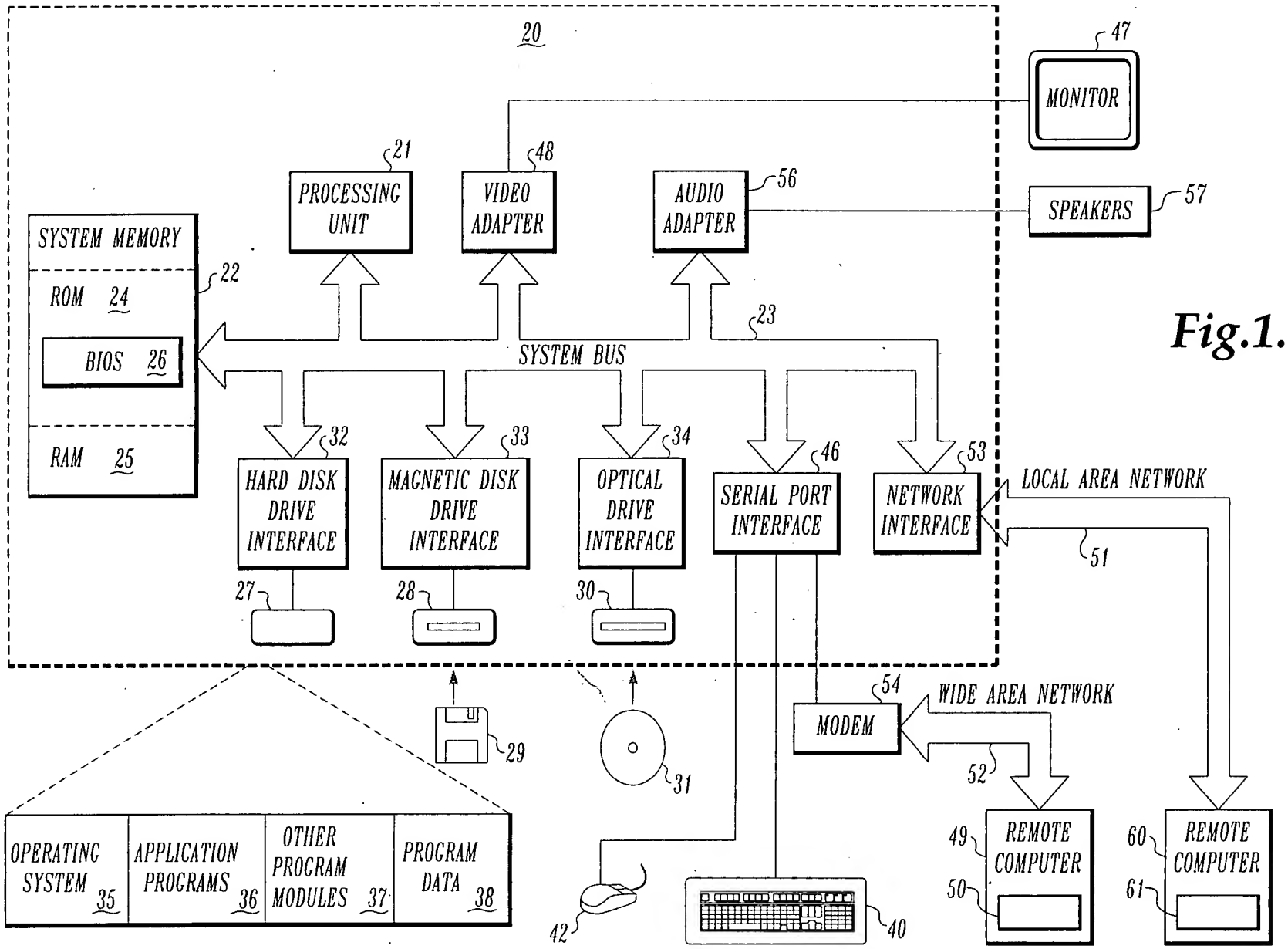

20
47
MONITOR
21
PROCESSING UNIT
48
VIDEO ADAPTER
56
AUDIO ADAPTER
SPEAKERS
57
SYSTEM MEMORY
22
ROM 24
BIOS 26
RAM 25
23
SYSTEM BUS
32
HARD DISK DRIVE INTERFACE
33
MAGNETIC DISK DRIVE INTERFACE
34
OPTICAL DRIVE INTERFACE
46
SERIAL PORT INTERFACE
53
NETWORK INTERFACE
LOCAL AREA NETWORK
51
27
28
30
29
31
54
MODEM
WIDE AREA NETWORK
52
OPERATING SYSTEM 35
APPLICATION PROGRAMS 36
OTHER PROGRAM MODULES 37
PROGRAM DATA 38
42
40
49
REMOTE COMPUTER
50
60
REMOTE COMPUTER
61

Fig.1.

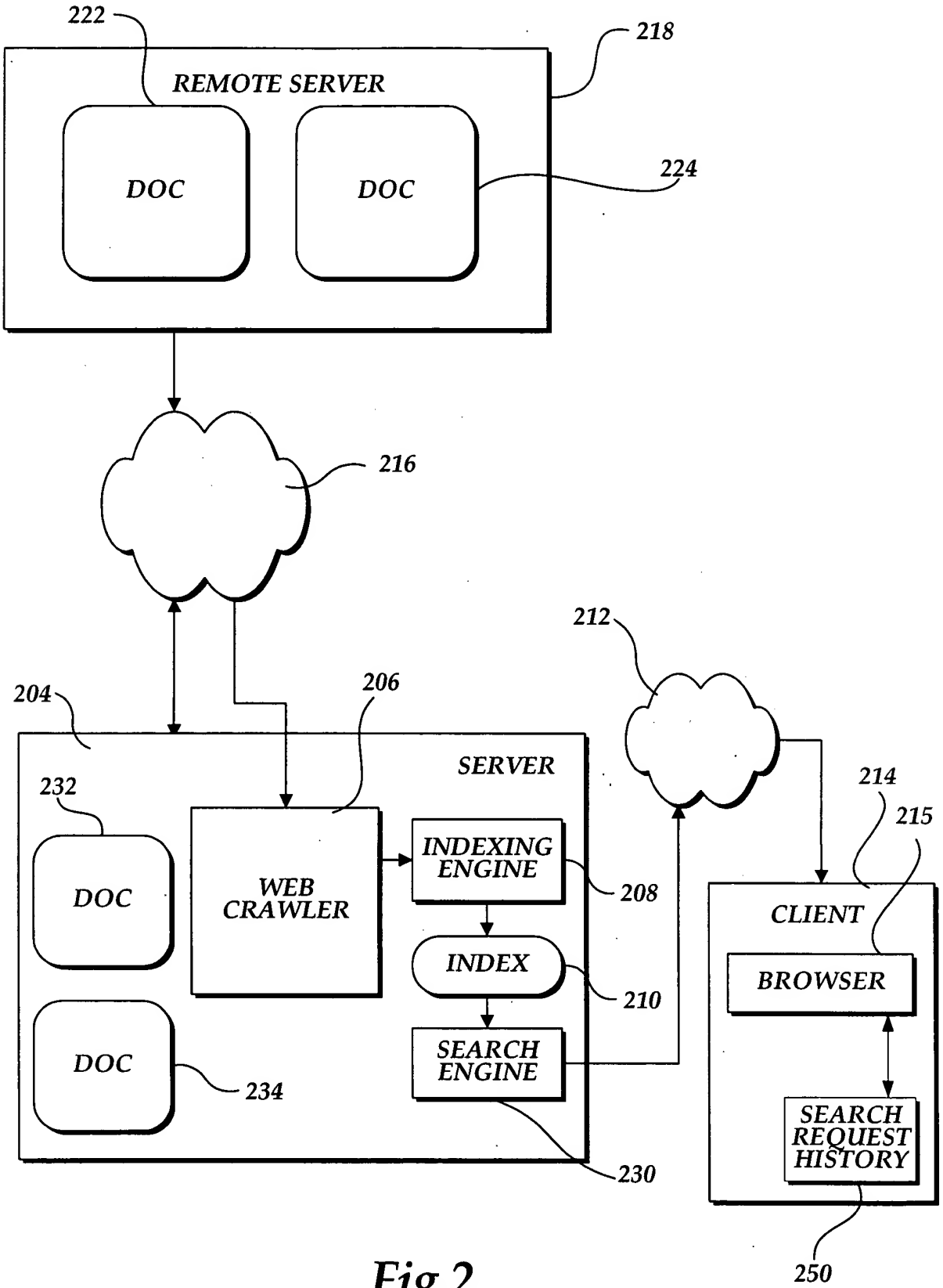

222
218
REMOTE SERVER
DOC
DOC
224
216
212
204
206
SERVER
232
DOC
WEB CRAWLER
INDEXING ENGINE
208
INDEX
210
SEARCH ENGINE
DOC
234
230
214
215
CLIENT
BROWSER
SEARCH REQUEST HISTORY
250

Fig.2.

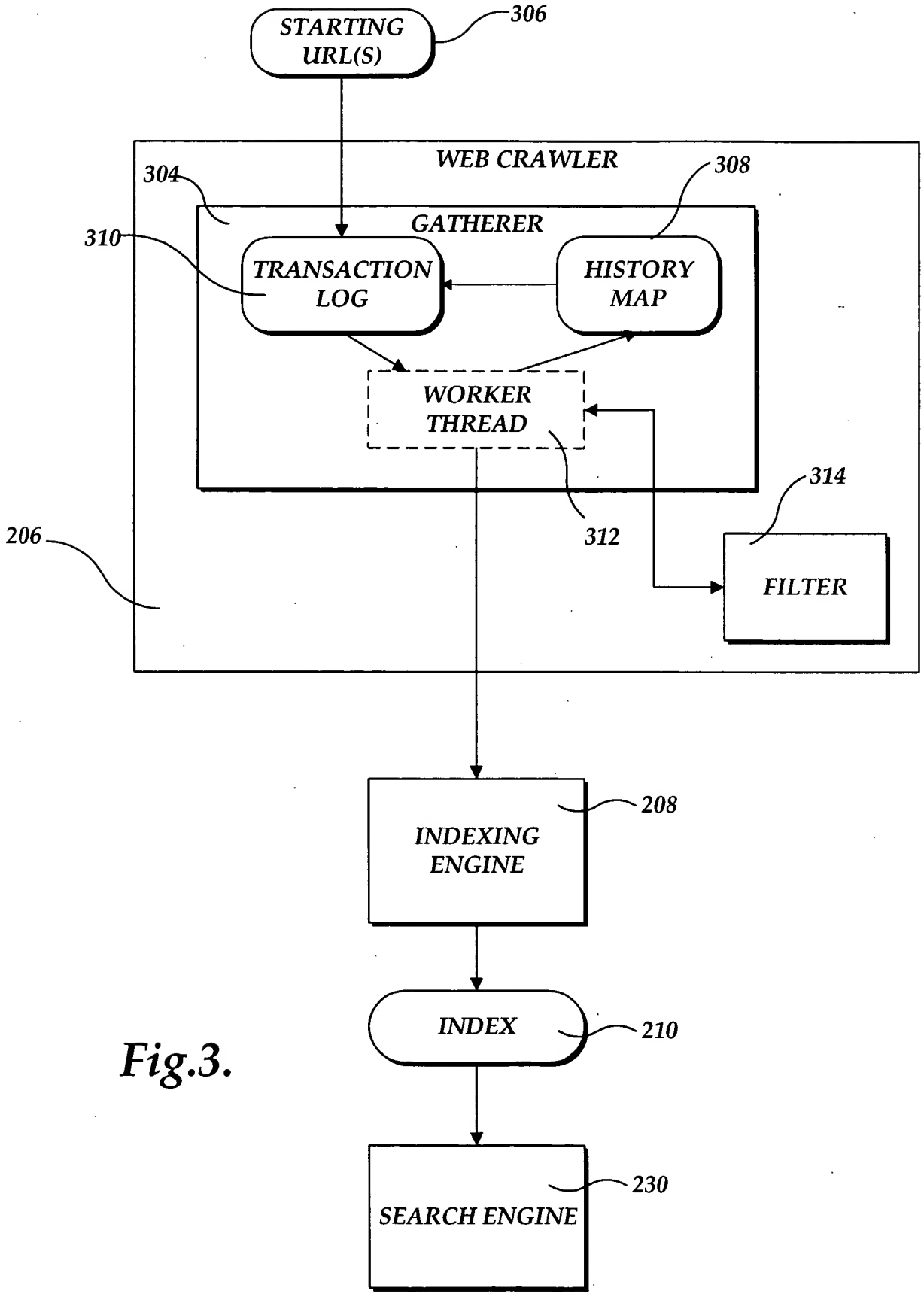
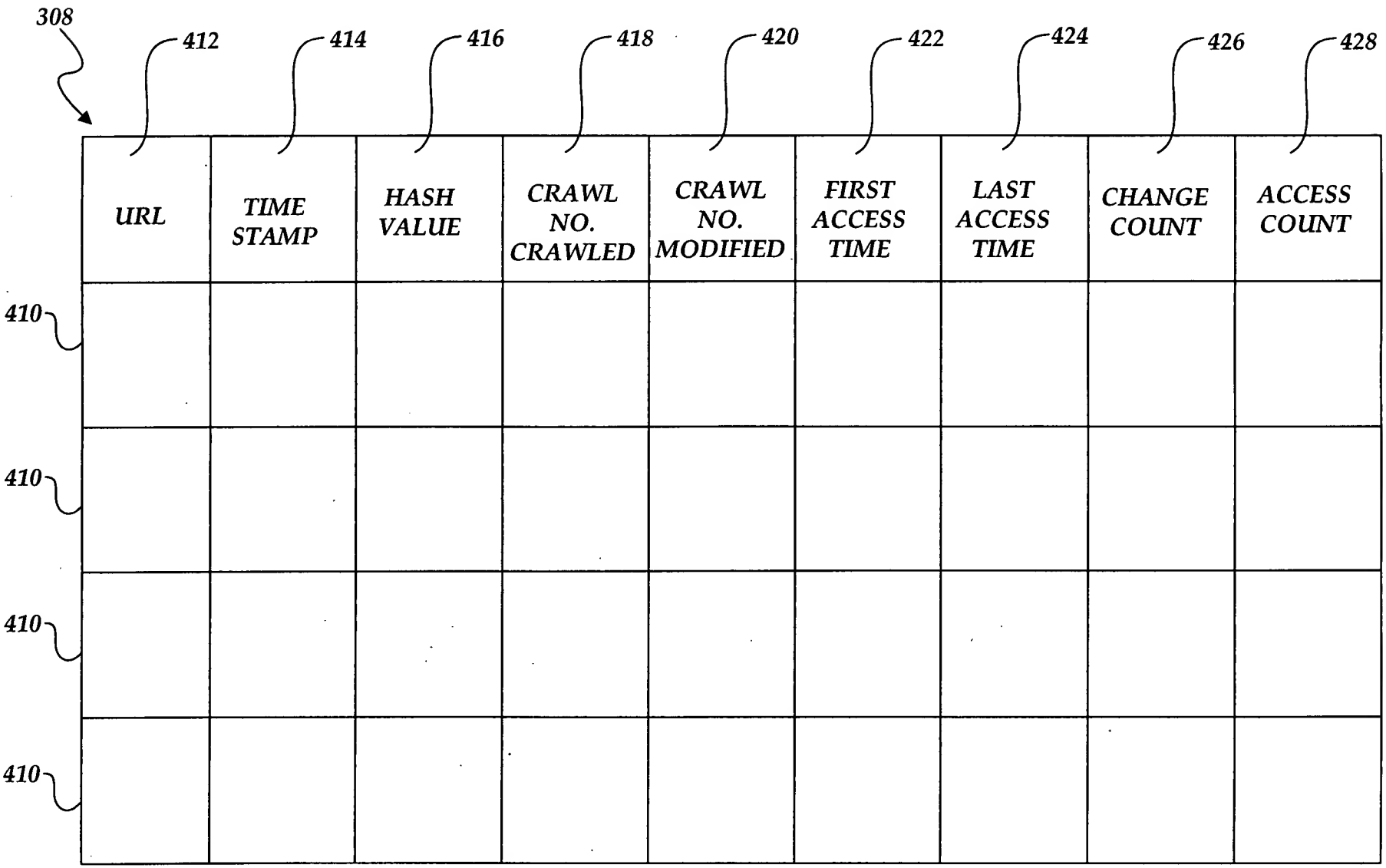

STARTING URL(S)
306
WEB CRAWLER
304
308
GATHERER
310
TRANSACTION LOG
HISTORY MAP
WORKER THREAD
312
314
206
FILTER
INDEXING ENGINE
208
INDEX
210
SEARCH ENGINE
230

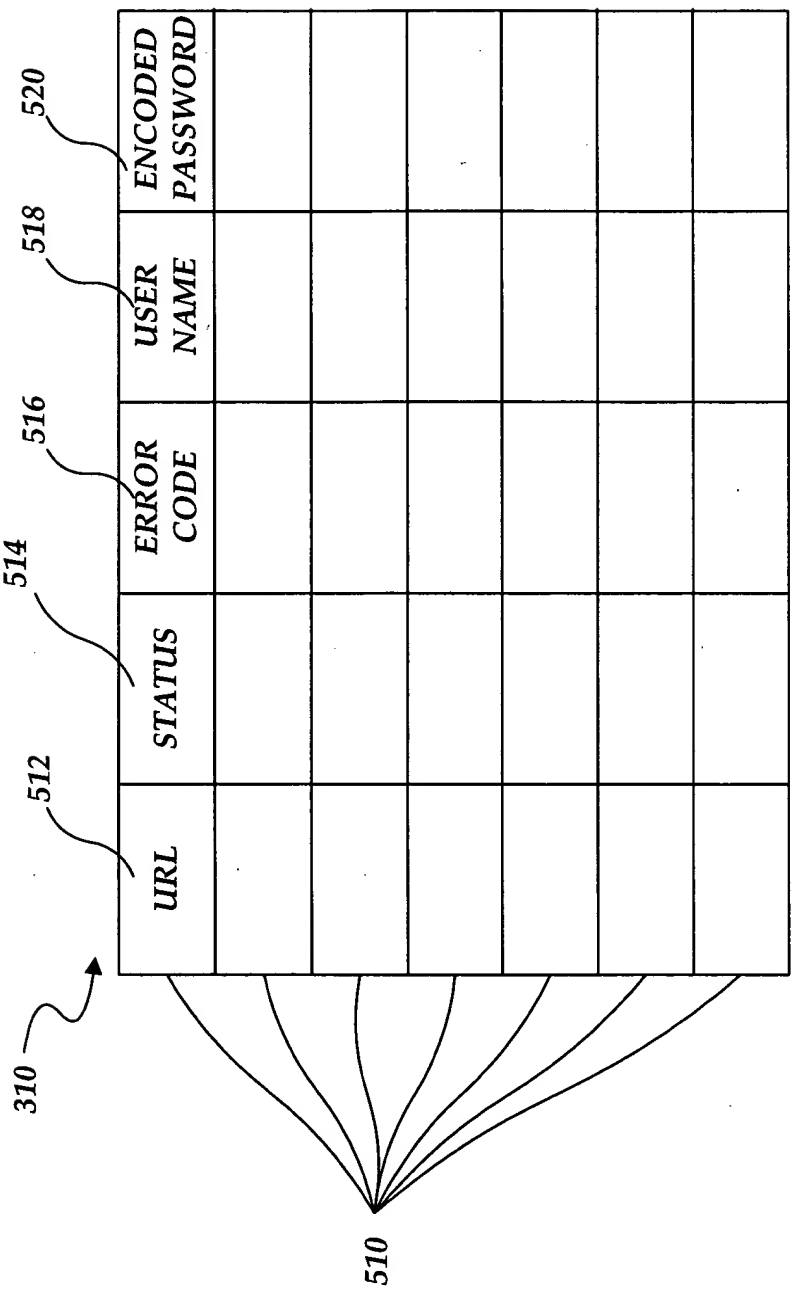
Fig.3.

0082T0" 8h7E6h60



HISTORY MAP

Fig.4.



TRANSACTION LOG

Fig.5.

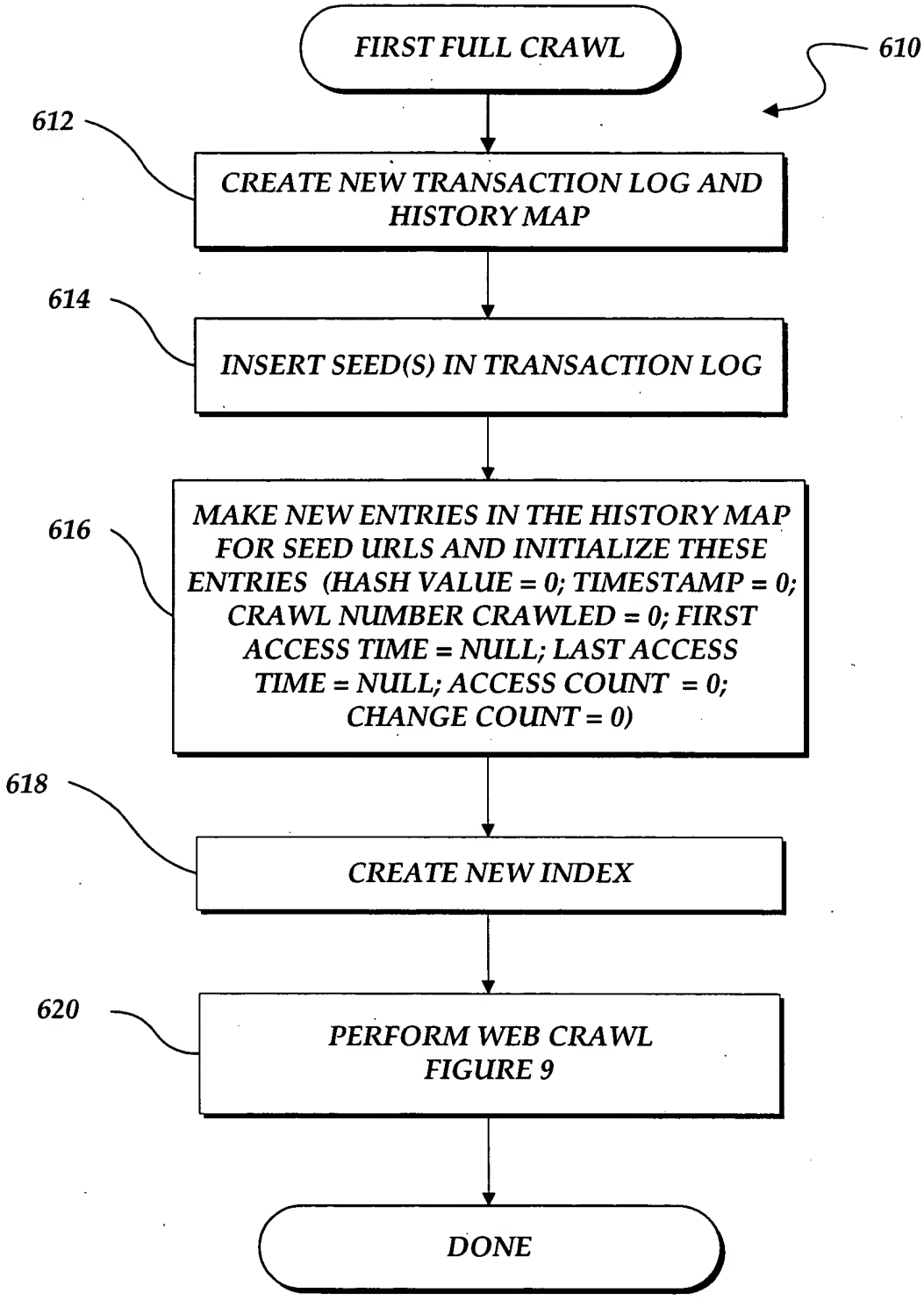


Fig.6.

09462760.8v7/6+60

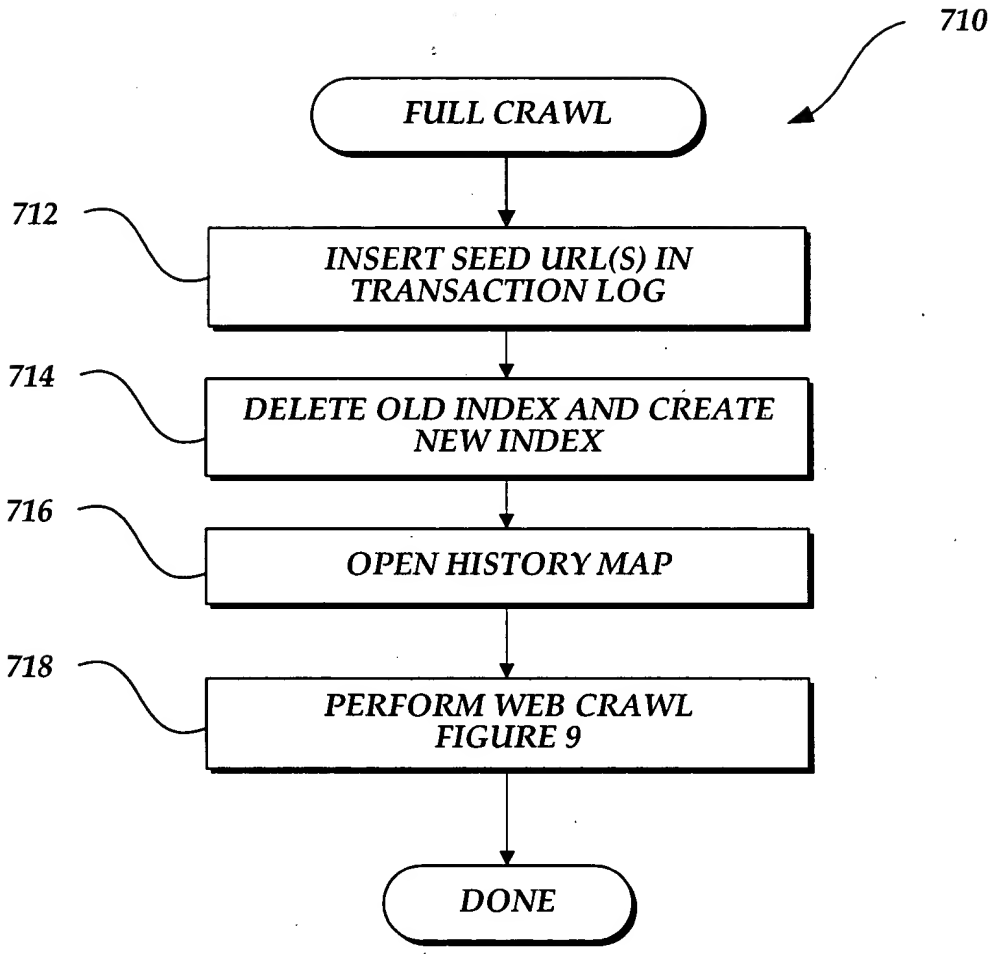


Fig.7.

0068170 84V6960

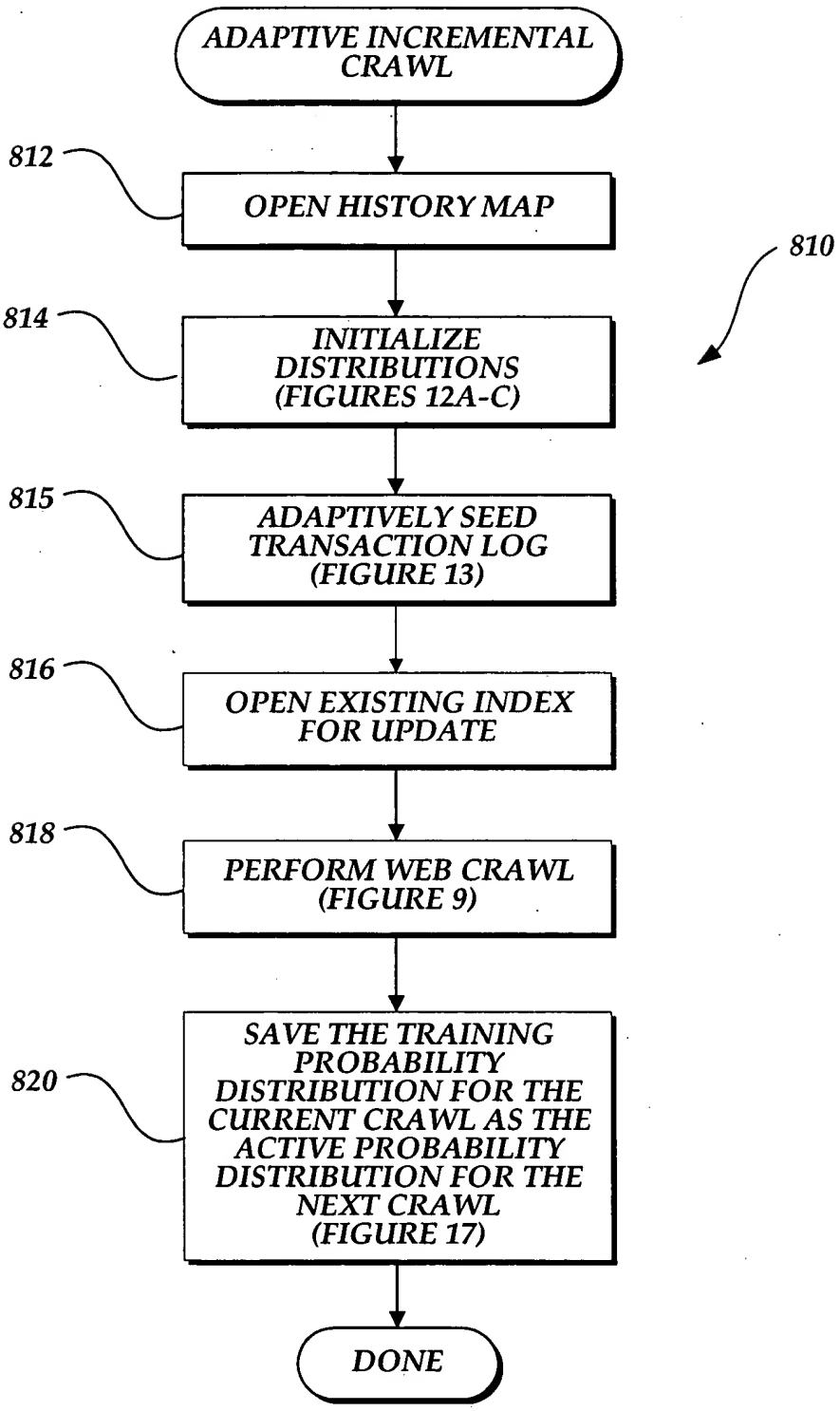

ADAPTIVE INCREMENTAL CRAWL
812
OPEN HISTORY MAP
810
814
INITIALIZE DISTRIBUTIONS (FIGURES 12A-C)
815
ADAPTIVELY SEED TRANSACTION LOG (FIGURE 13)
816
OPEN EXISTING INDEX FOR UPDATE
818
PERFORM WEB CRAWL (FIGURE 9)
820
SAVE THE TRAINING PROBABILITY DISTRIBUTION FOR THE CURRENT CRAWL AS THE ACTIVE PROBABILITY DISTRIBUTION FOR THE NEXT CRAWL (FIGURE 17)
DONE

Fig.8.

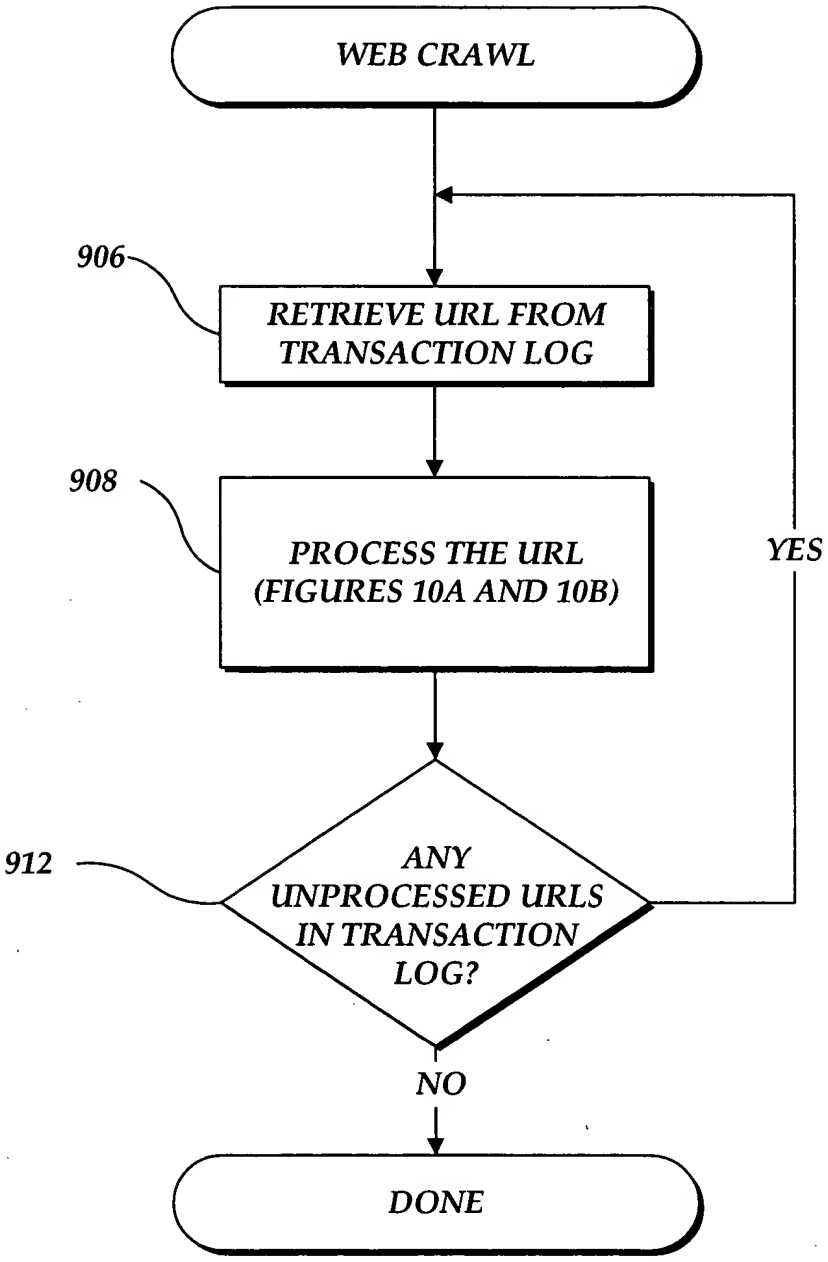

WEB CRAWL
906
RETRIEVE URL FROM TRANSACTION LOG
908
PROCESS THE URL (FIGURES 10A AND 10B)
912
ANY UNPROCESSED URLS IN TRANSACTION LOG?
YES
NO
DONE

Fig.9.

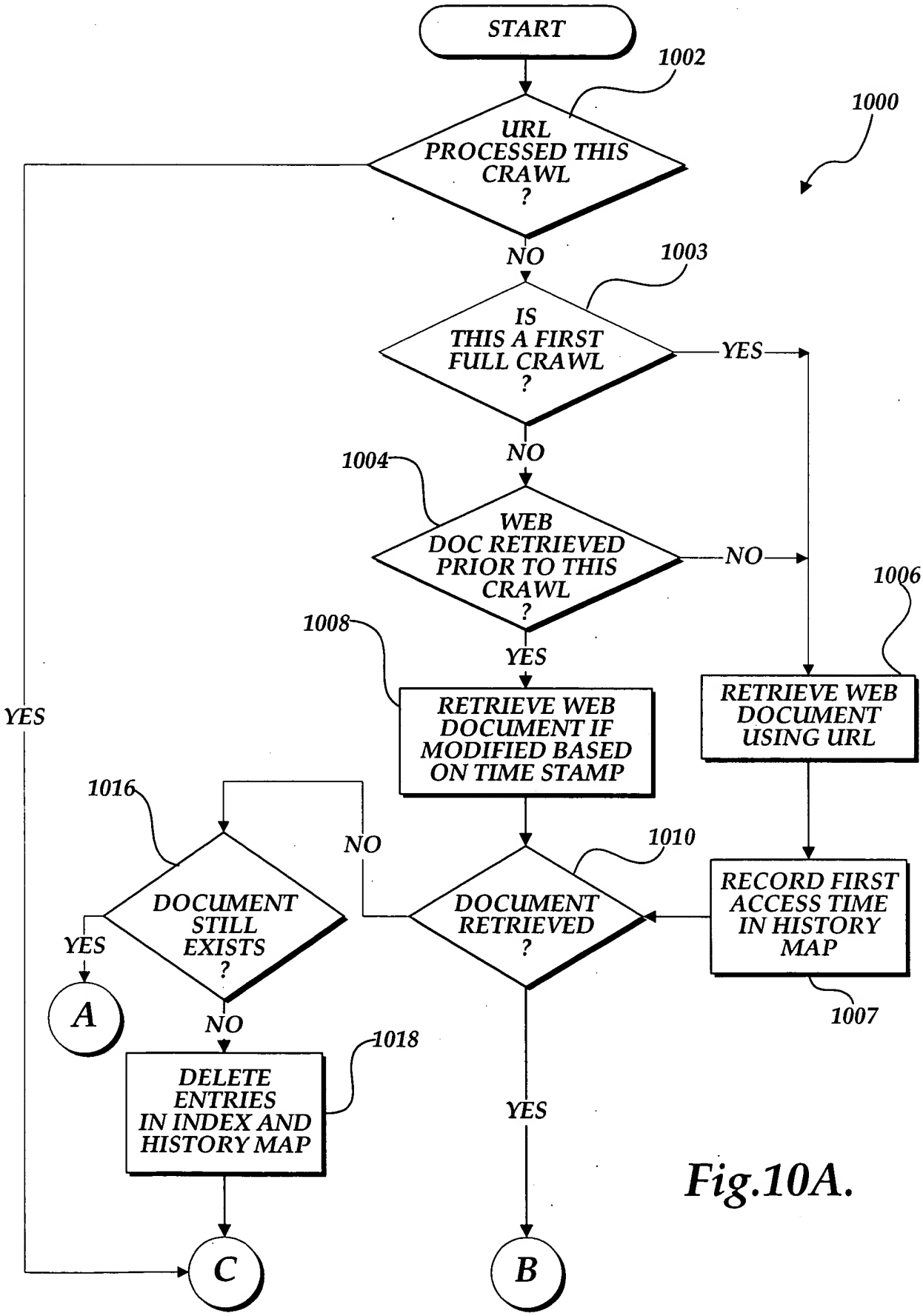

START
1002
1000
URL PROCESSED THIS CRAWL ?
NO
1003
IS THIS A FIRST FULL CRAWL ?
YES
NO
1004
WEB DOC RETRIEVED PRIOR TO THIS CRAWL ?
NO
1006
YES
1008
RETRIEVE WEB DOCUMENT IF MODIFIED BASED ON TIME STAMP
RETRIEVE WEB DOCUMENT USING URL
YES
1016
NO
1010
DOCUMENT STILL EXISTS ?
DOCUMENT RETRIEVED ?
RECORD FIRST ACCESS TIME IN HISTORY MAP
1007
YES
A
NO
1018
DELETE ENTRIES IN INDEX AND HISTORY MAP
YES
C
B

Fig.10A.

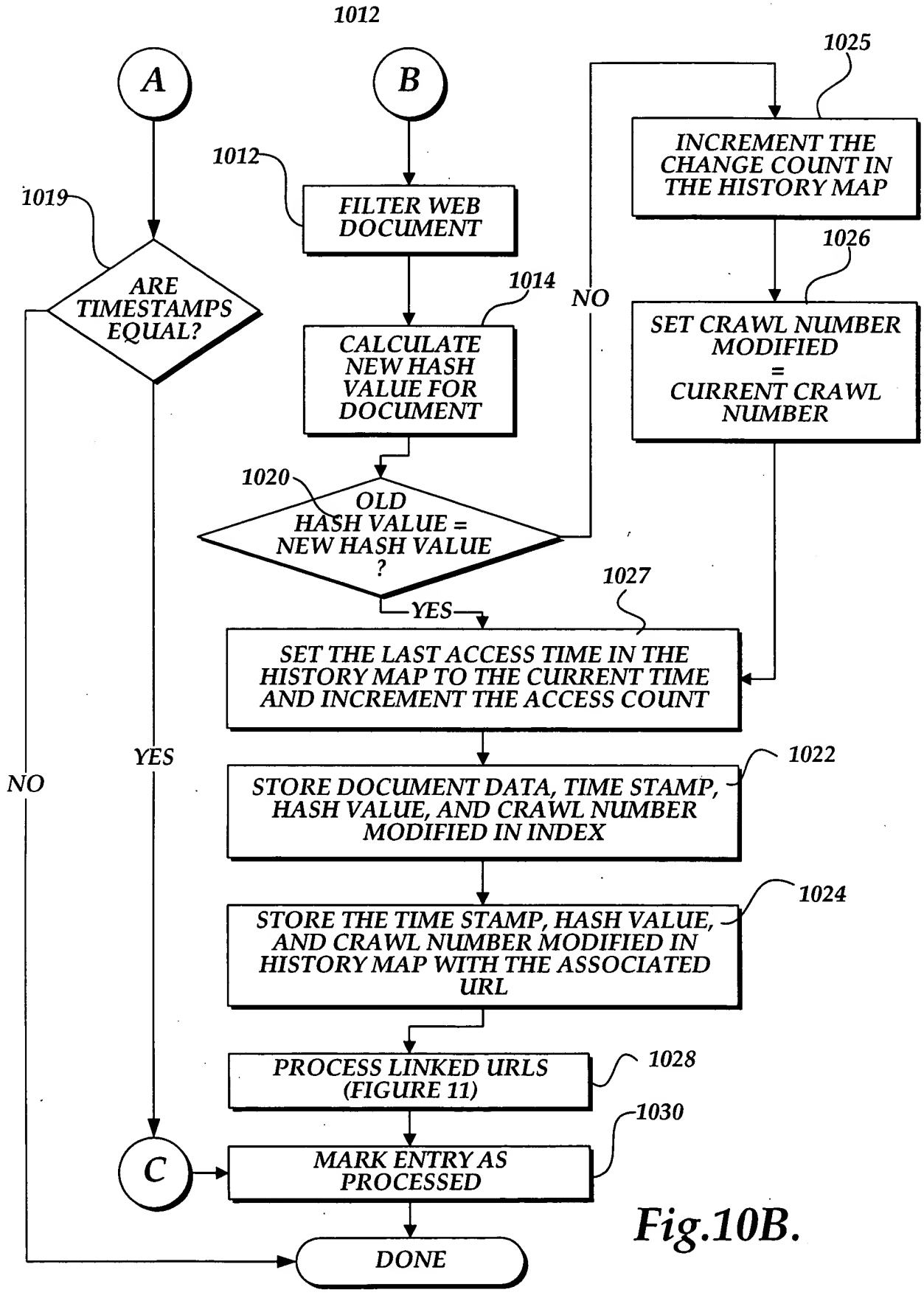

1012
A
B
1025
INCREMENT THE CHANGE COUNT IN THE HISTORY MAP
1012
FILTER WEB DOCUMENT
1019
ARE TIMESTAMPS EQUAL?
1014
NO
1026
SET CRAWL NUMBER MODIFIED = CURRENT CRAWL NUMBER
CALCULATE NEW HASH VALUE FOR DOCUMENT
1020
OLD HASH VALUE = NEW HASH VALUE ?
1027
YES
SET THE LAST ACCESS TIME IN THE HISTORY MAP TO THE CURRENT TIME AND INCREMENT THE ACCESS COUNT
YES
NO
1022
STORE DOCUMENT DATA, TIME STAMP, HASH VALUE, AND CRAWL NUMBER MODIFIED IN INDEX
1024
STORE THE TIME STAMP, HASH VALUE, AND CRAWL NUMBER MODIFIED IN HISTORY MAP WITH THE ASSOCIATED URL
PROCESS LINKED URLS (FIGURE 11)
1028
1030
C
MARK ENTRY AS PROCESSED
DONE

Fig.10B.

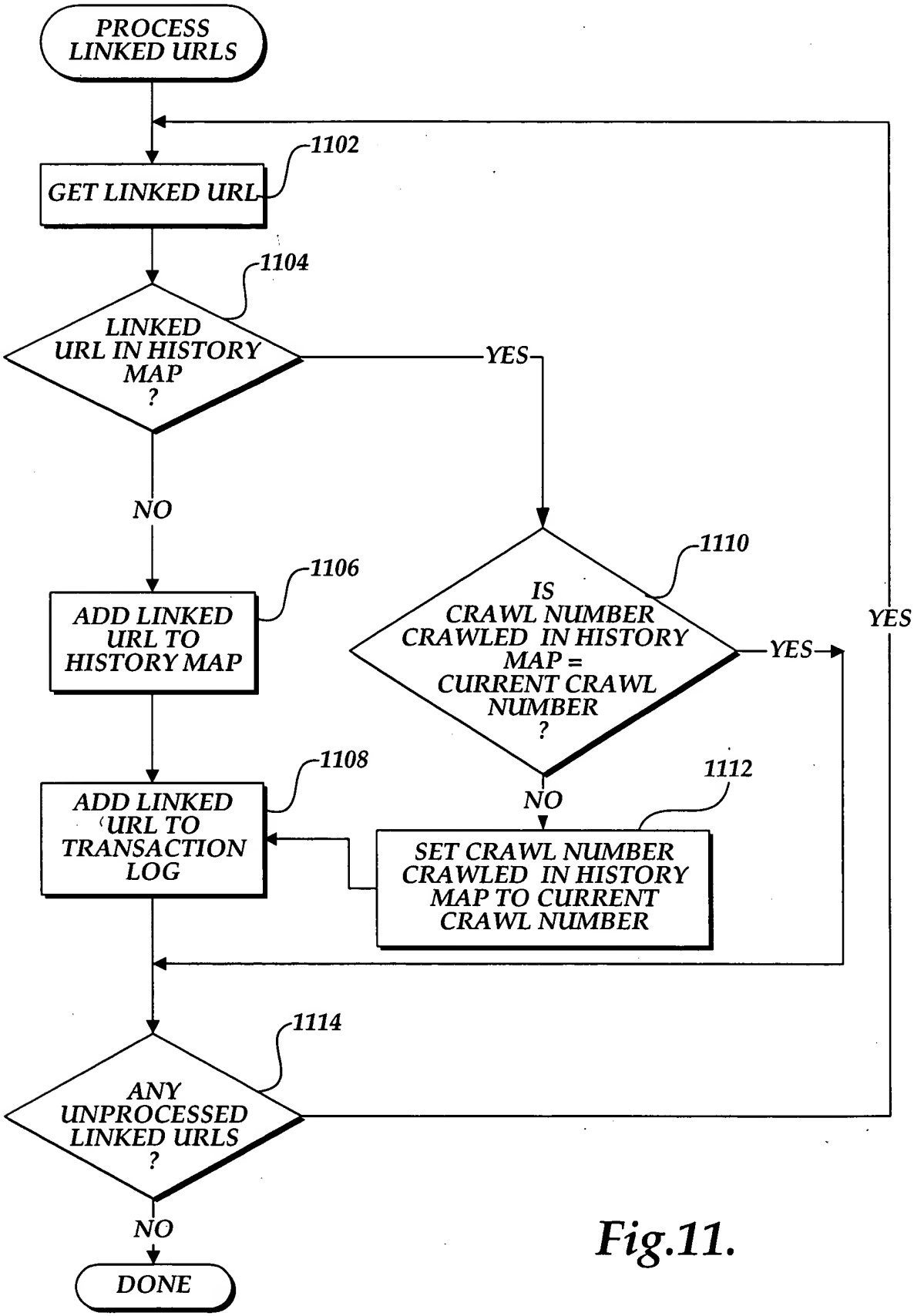

PROCESS LINKED URLS
1102
GET LINKED URL
1104
LINKED URL IN HISTORY MAP ?
YES
NO
1106
ADD LINKED URL TO HISTORY MAP
1110
IS CRAWL NUMBER CRAWLED IN HISTORY MAP = CURRENT CRAWL NUMBER ?
YES
YES
NO
1108
ADD LINKED URL TO TRANSACTION LOG
1112
SET CRAWL NUMBER CRAWLED IN HISTORY MAP TO CURRENT CRAWL NUMBER
1114
ANY UNPROCESSED LINKED URLS ?
NO
DONE

Fig.11.

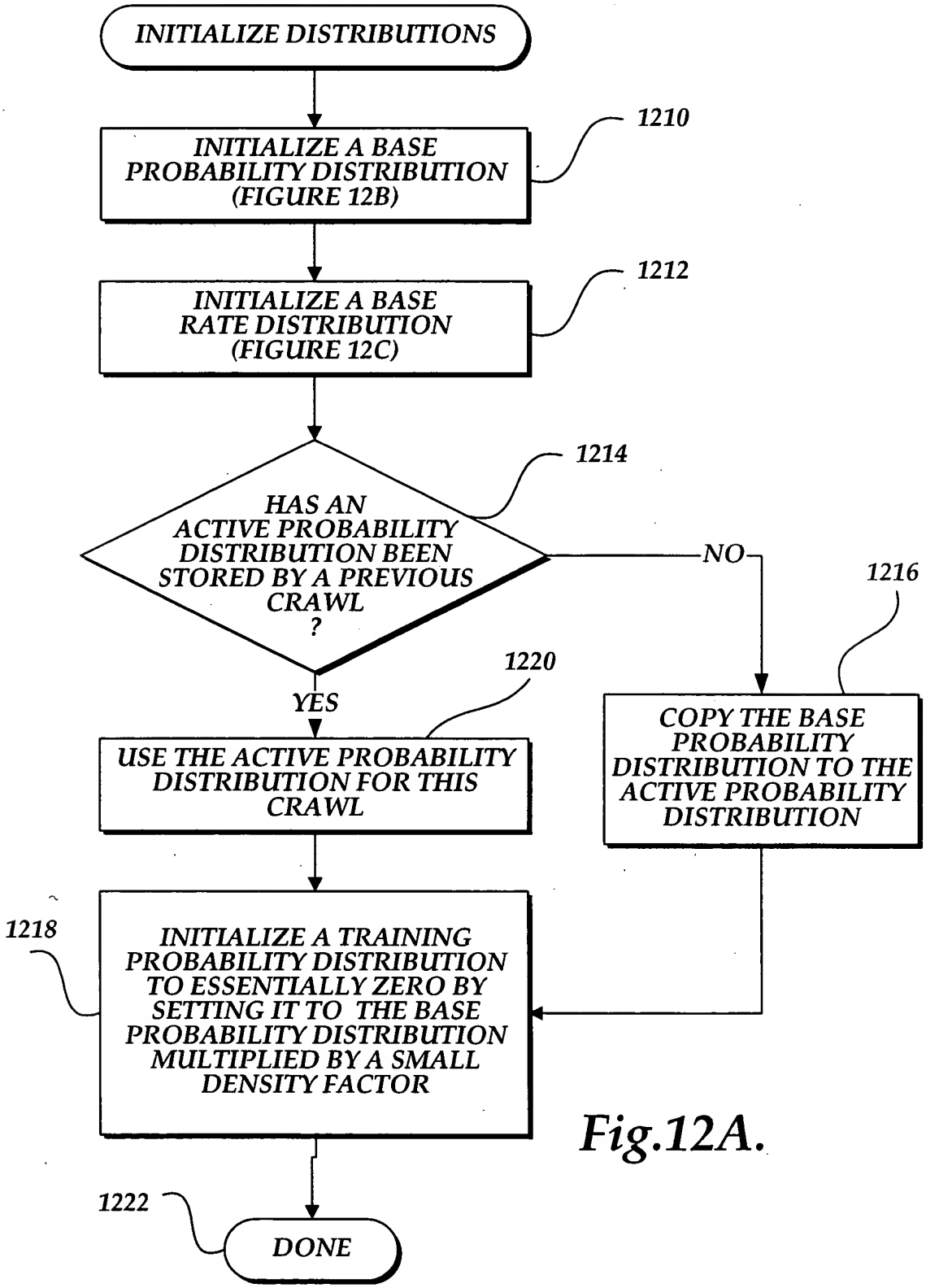

INITIALIZE DISTRIBUTIONS
INITIALIZE A BASE PROBABILITY DISTRIBUTION (FIGURE 12B)
1210
INITIALIZE A BASE RATE DISTRIBUTION (FIGURE 12C)
1212
HAS AN ACTIVE PROBABILITY DISTRIBUTION BEEN STORED BY A PREVIOUS CRAWL ?
1214
NO
YES
1216
COPY THE BASE PROBABILITY DISTRIBUTION TO THE ACTIVE PROBABILITY DISTRIBUTION
1220
USE THE ACTIVE PROBABILITY DISTRIBUTION FOR THIS CRAWL
1218
INITIALIZE A TRAINING PROBABILITY DISTRIBUTION TO ESSENTIALLY ZERO BY SETTING IT TO THE BASE PROBABILITY DISTRIBUTION MULTIPLIED BY A SMALL DENSITY FACTOR
1222
DONE

Fig.12A.

0082TO" 8h7E6h60

BASE PROBABILITY DISTRIBUTION

SAMPLE PROBABILITIES (N)	BASE PROBABILITY OF DOCUMENT CHANGE DURING INTERVAL (N)
P1	.3/(n-1)
P2	.3/(n-1)
P3	.3/(n-1)
P4	.3/(n-1)
P5	.3/(n-1)
P6	.3/(n-1)
P7	.3/(n-1)
P8	.3/(n-1)
P9	.3/(n-1)
Pn-1	.3/(n-1)
Pn	.7

30% OF DOCUMENTS ARE ASSUMED TO CHANGE AND ARE UNIFORMLY DISTRIBUTED AMONG THE SAMPLE CHANGE RATES . THE NUMBER OF DOCUMENTS WITH A GIVEN CHANGE RATE IS EXPRESSED AS A PERCENTAGE OF ALL PREVIOUSLY RETRIEVED DOCUMENTS.

70% OF DOCUMENTS ASSUMED TO NEVER CHANGE

Fig.12B.

00820 "8h7E6h60

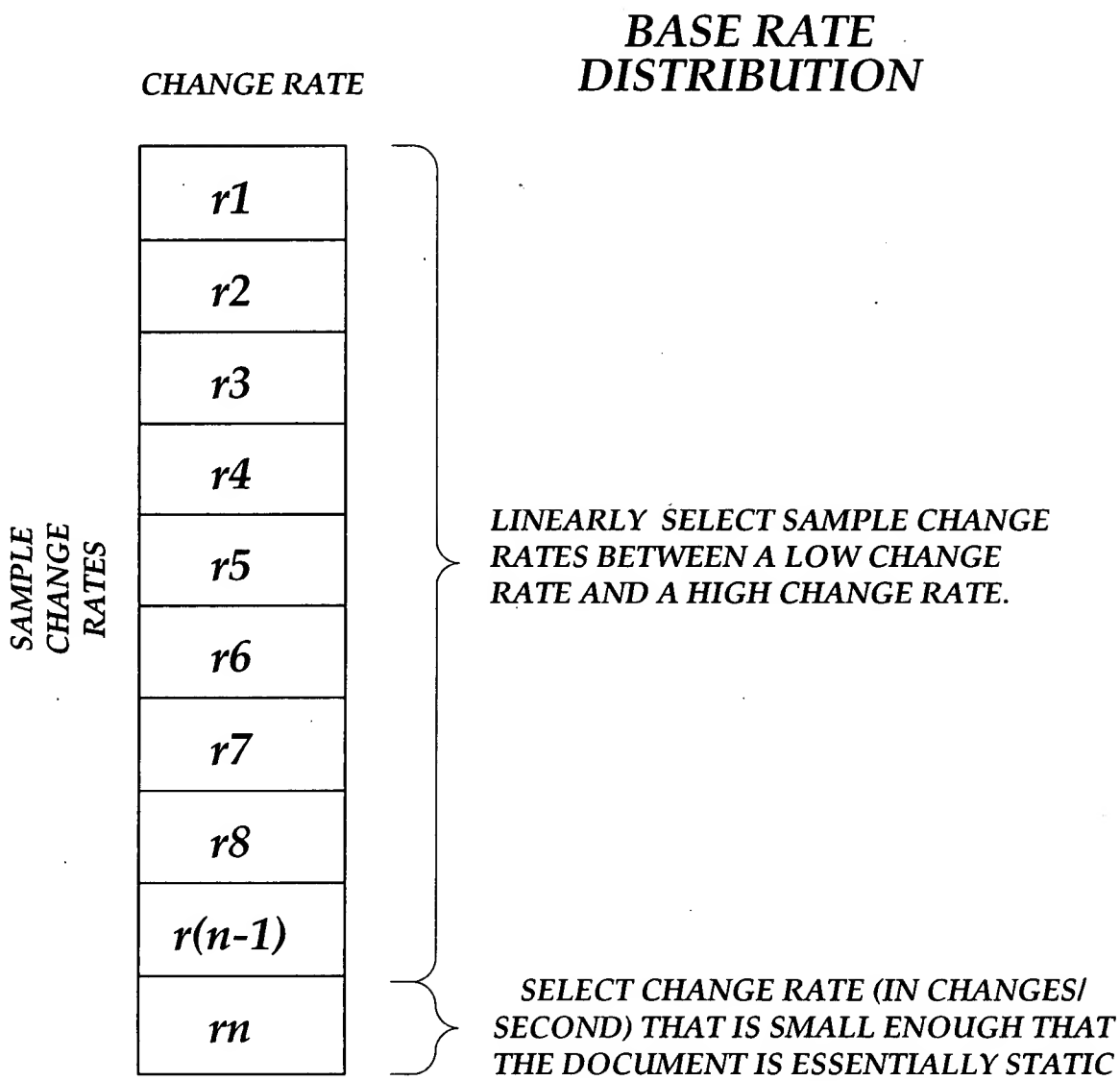


Fig.12C.

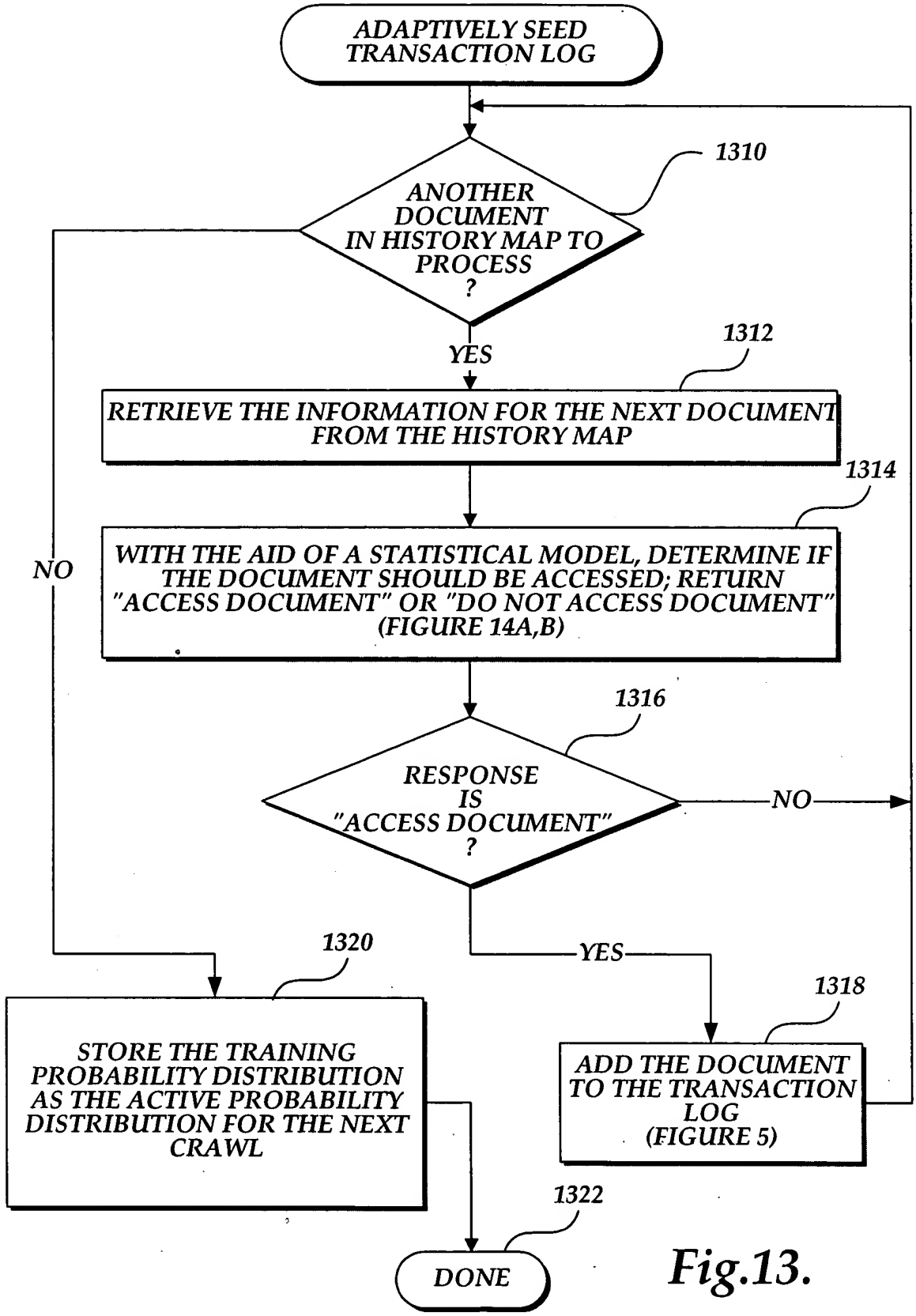


Fig.13.

09482760 - 84762760

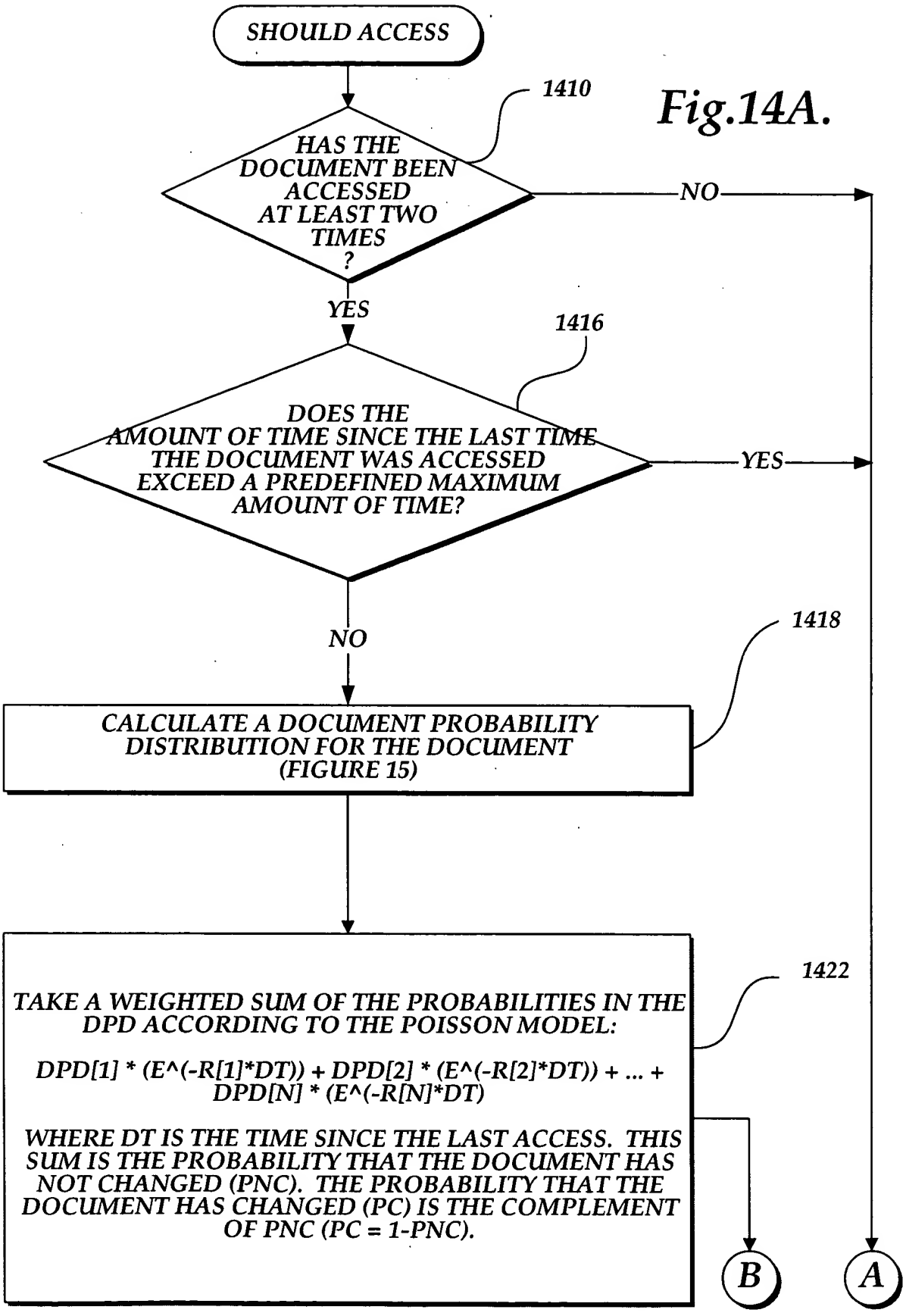


Fig.14A.

0098210 - 8472676D

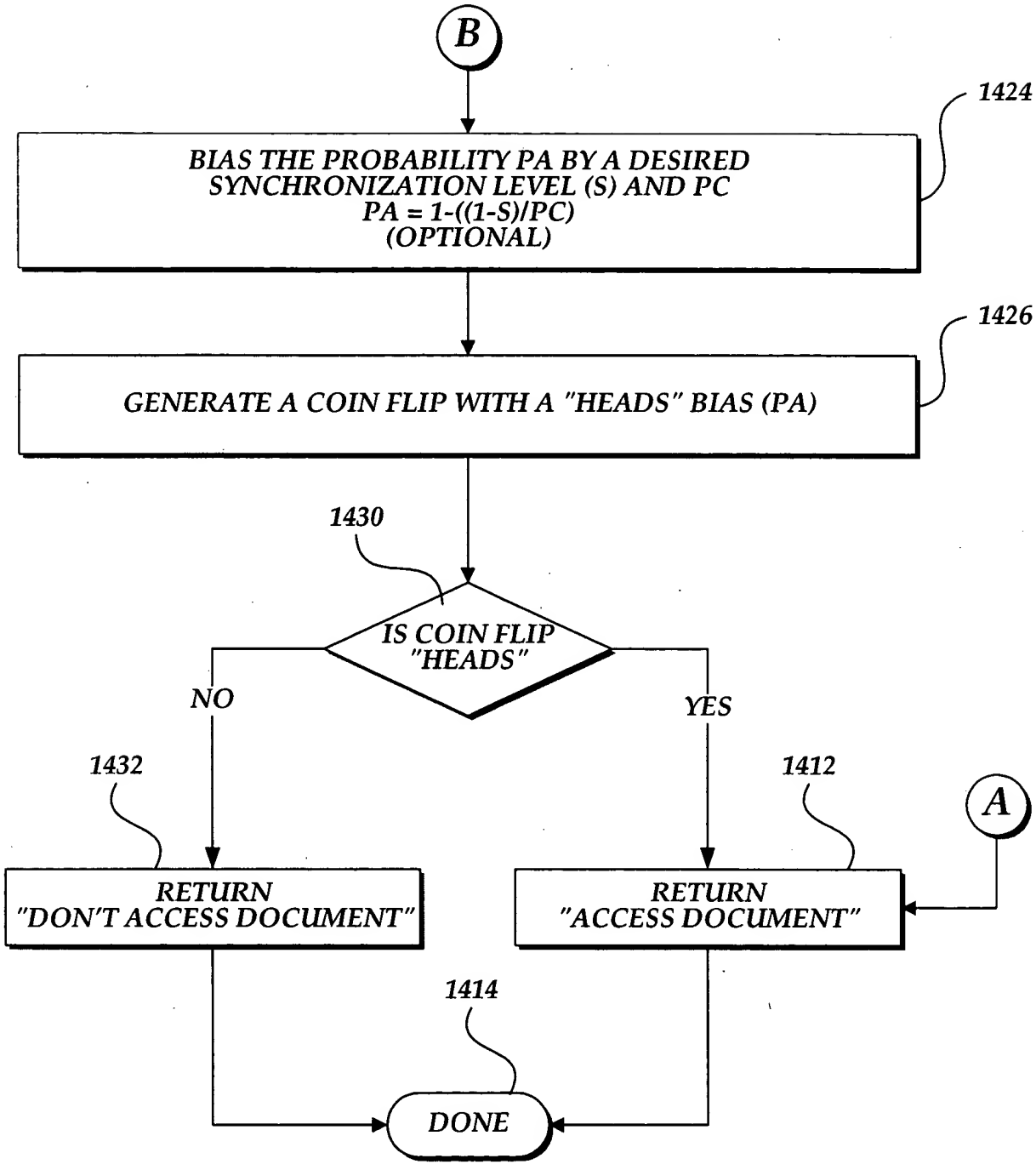


Fig.14B.

09593748 - 042800

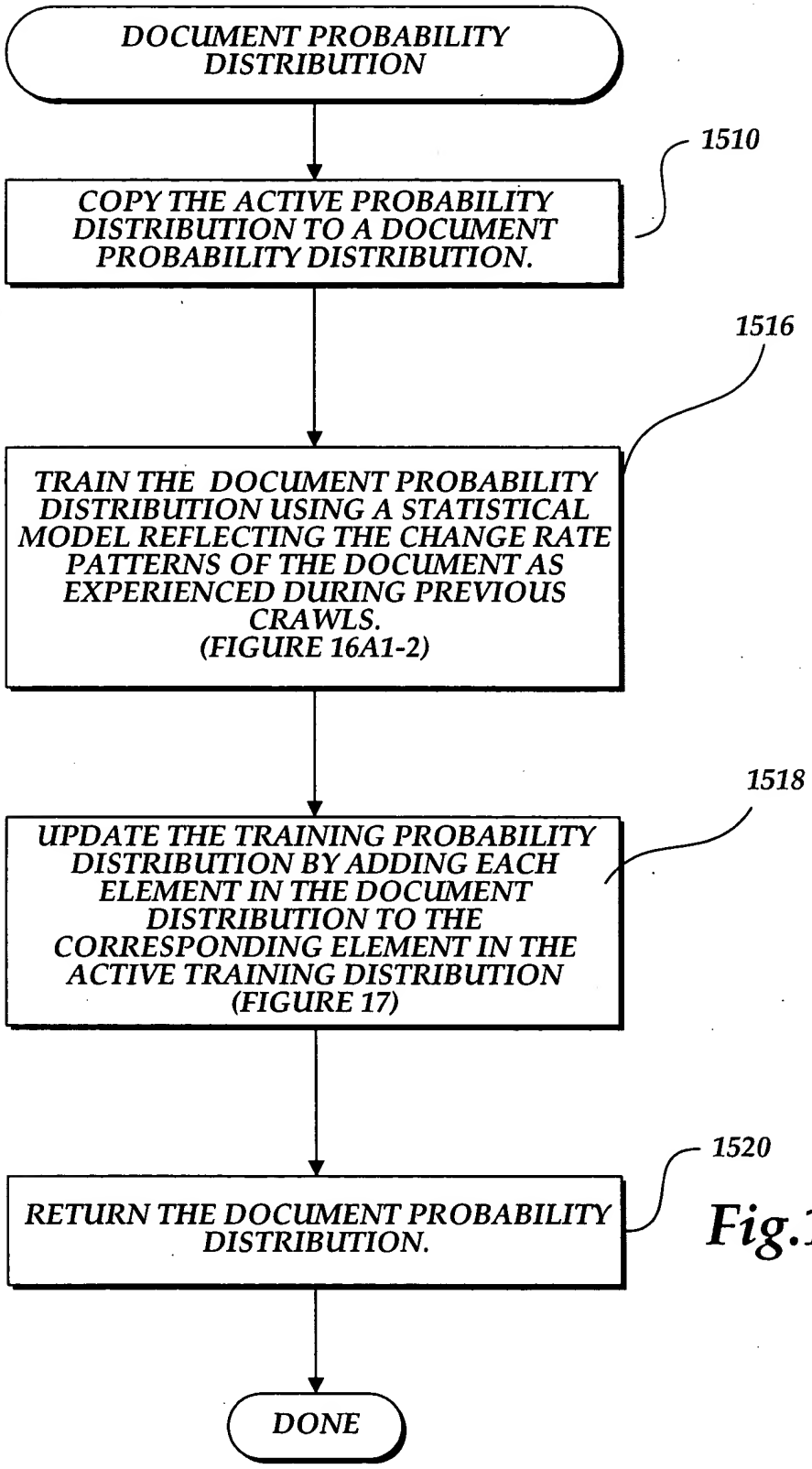


Fig.15.

0098270 - 0942680

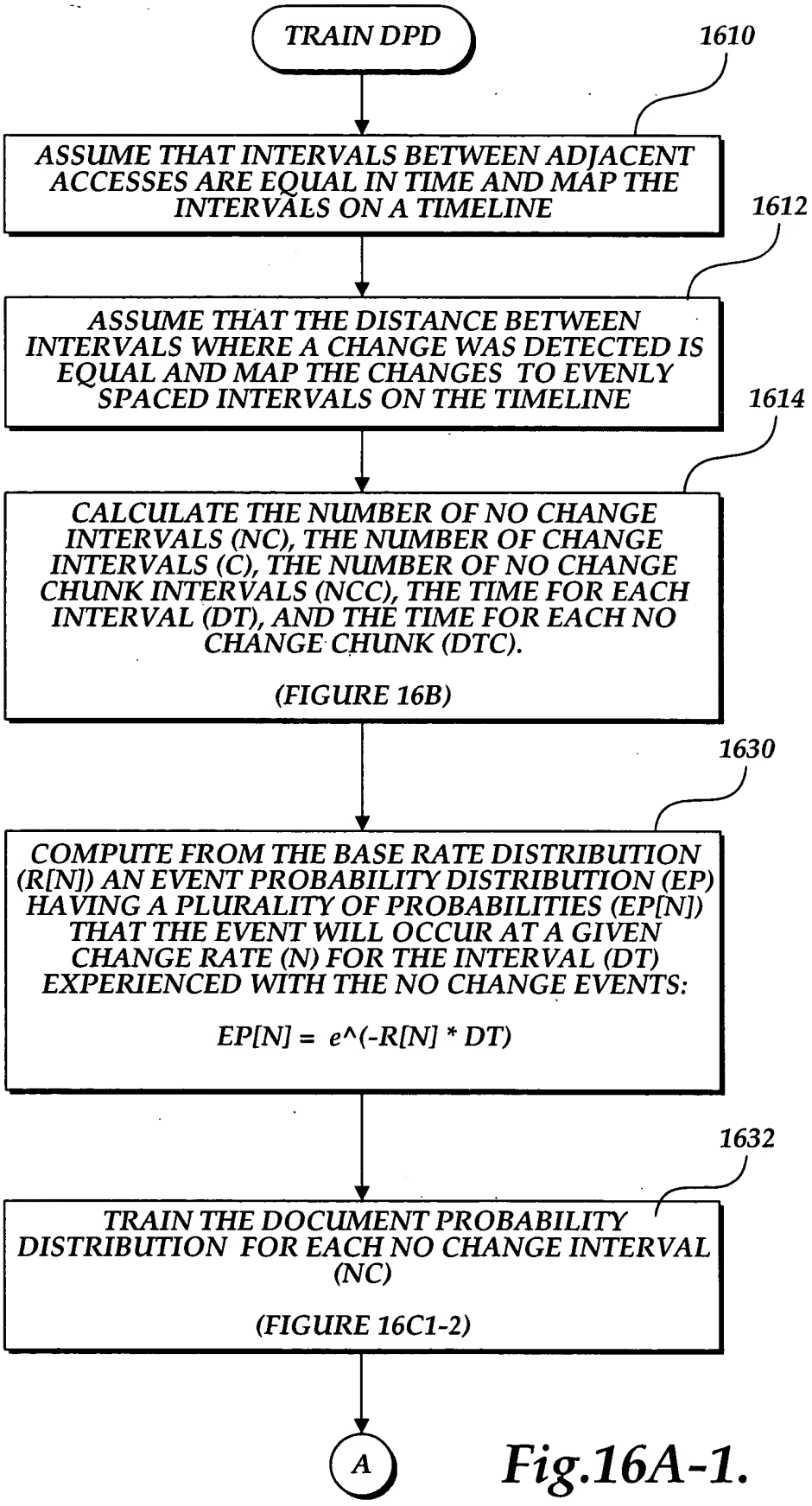

TRAIN DPD
1610
ASSUME THAT INTERVALS BETWEEN ADJACENT ACCESSES ARE EQUAL IN TIME AND MAP THE INTERVALS ON A TIMELINE
1612
ASSUME THAT THE DISTANCE BETWEEN INTERVALS WHERE A CHANGE WAS DETECTED IS EQUAL AND MAP THE CHANGES TO EVENLY SPACED INTERVALS ON THE TIMELINE
1614
CALCULATE THE NUMBER OF NO CHANGE INTERVALS (NC), THE NUMBER OF CHANGE INTERVALS (C), THE NUMBER OF NO CHANGE CHUNK INTERVALS (NCC), THE TIME FOR EACH INTERVAL (DT), AND THE TIME FOR EACH NO CHANGE CHUNK (DTC).
(FIGURE 16B)
1630
COMPUTE FROM THE BASE RATE DISTRIBUTION (R[N]) AN EVENT PROBABILITY DISTRIBUTION (EP) HAVING A PLURALITY OF PROBABILITIES (EP[N]) THAT THE EVENT WILL OCCUR AT A GIVEN CHANGE RATE (N) FOR THE INTERVAL (DT) EXPERIENCED WITH THE NO CHANGE EVENTS:
EP[N] = e^(-R[N] * DT)
1632
TRAIN THE DOCUMENT PROBABILITY DISTRIBUTION FOR EACH NO CHANGE INTERVAL (NC)
(FIGURE 16C1-2)
A

Fig.16A-1.

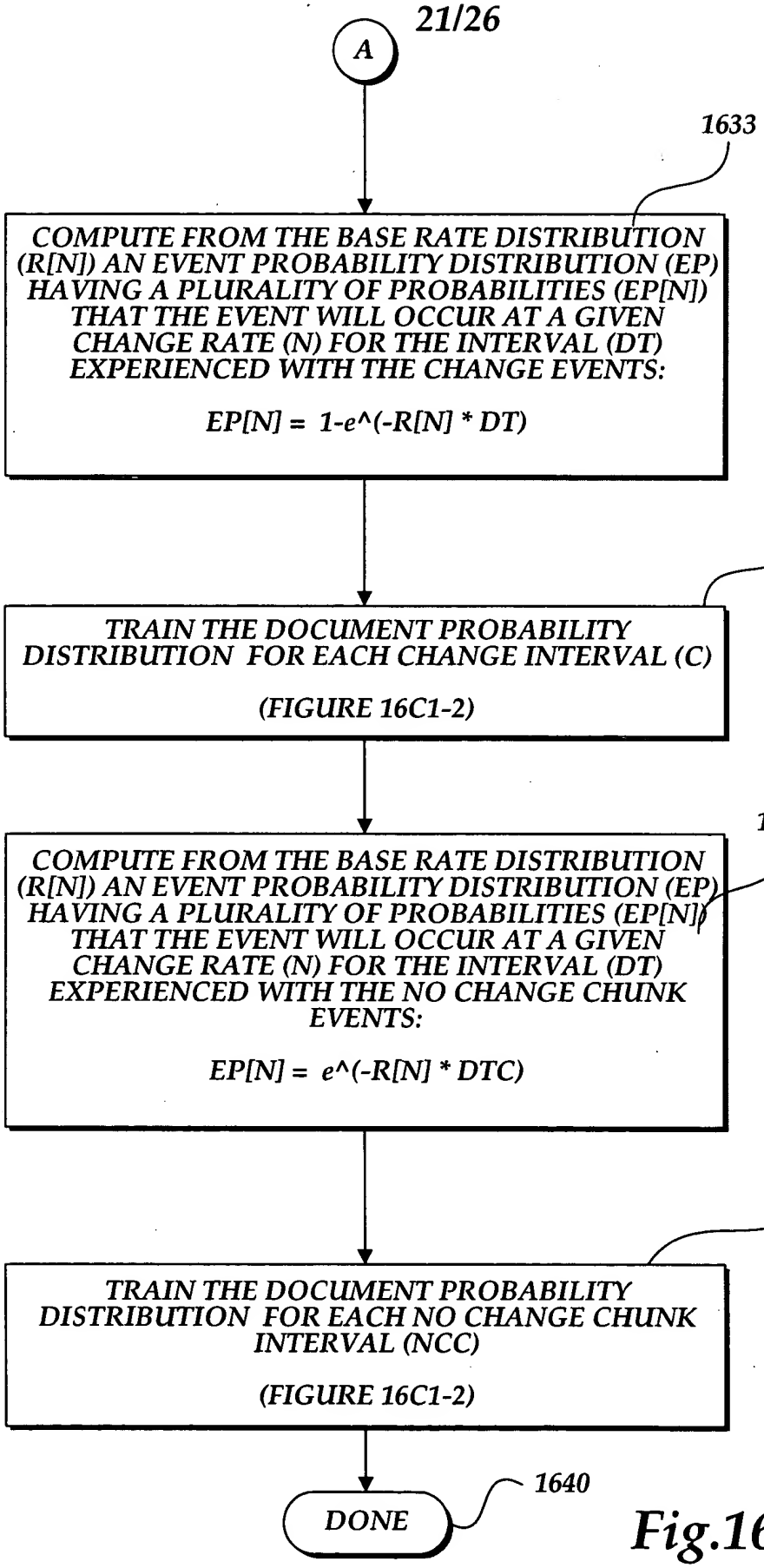


Fig.16A-2.

0094-9602-0-6h-36-pdf

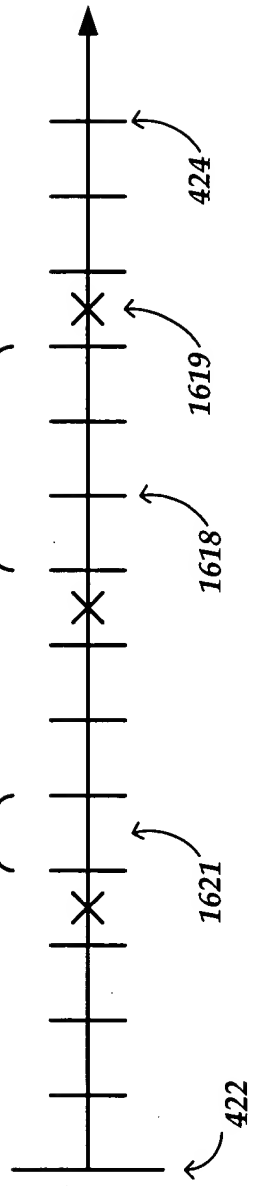


FIG.16B.

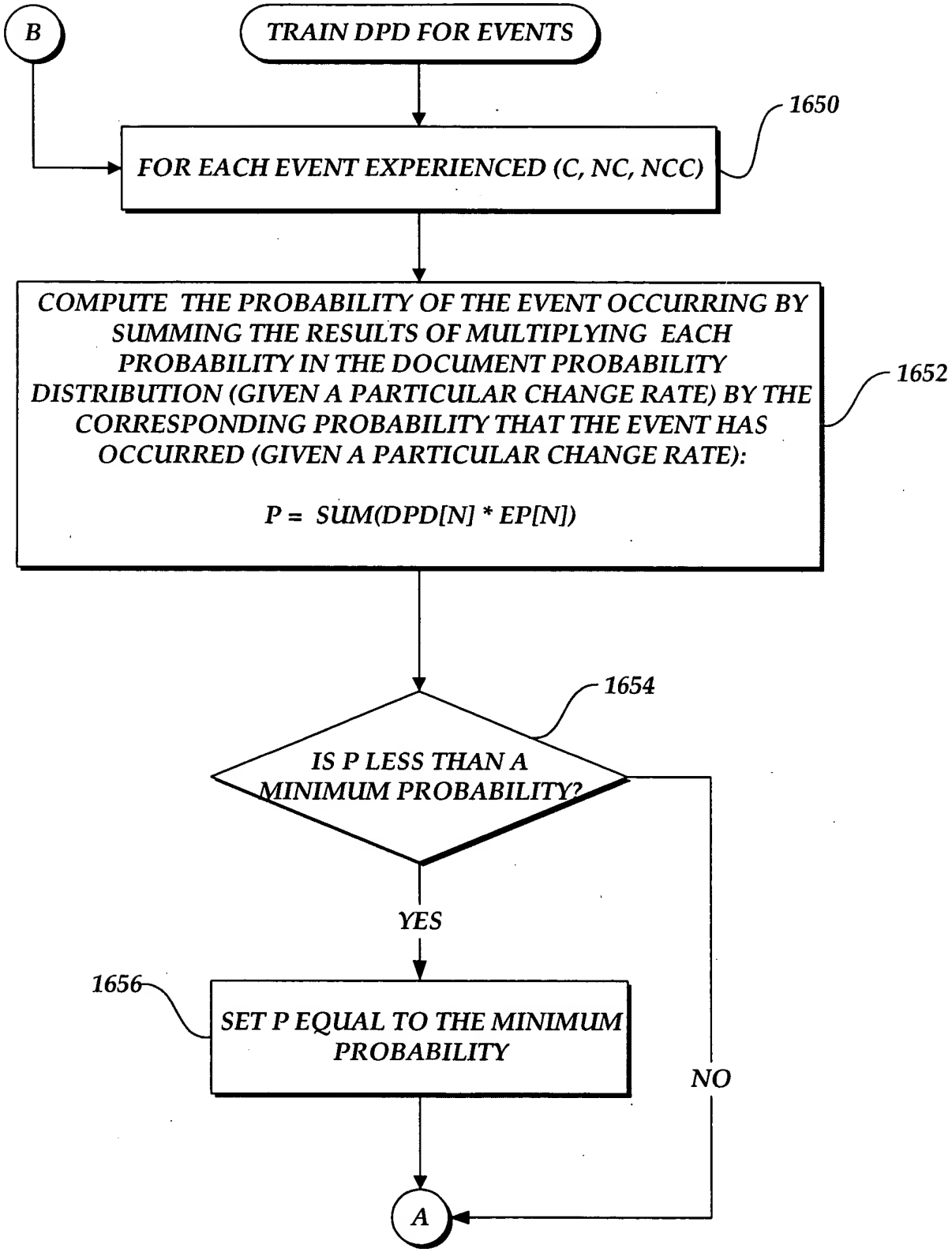


Fig.16C-1.

09/45/2004 - 08:21:00

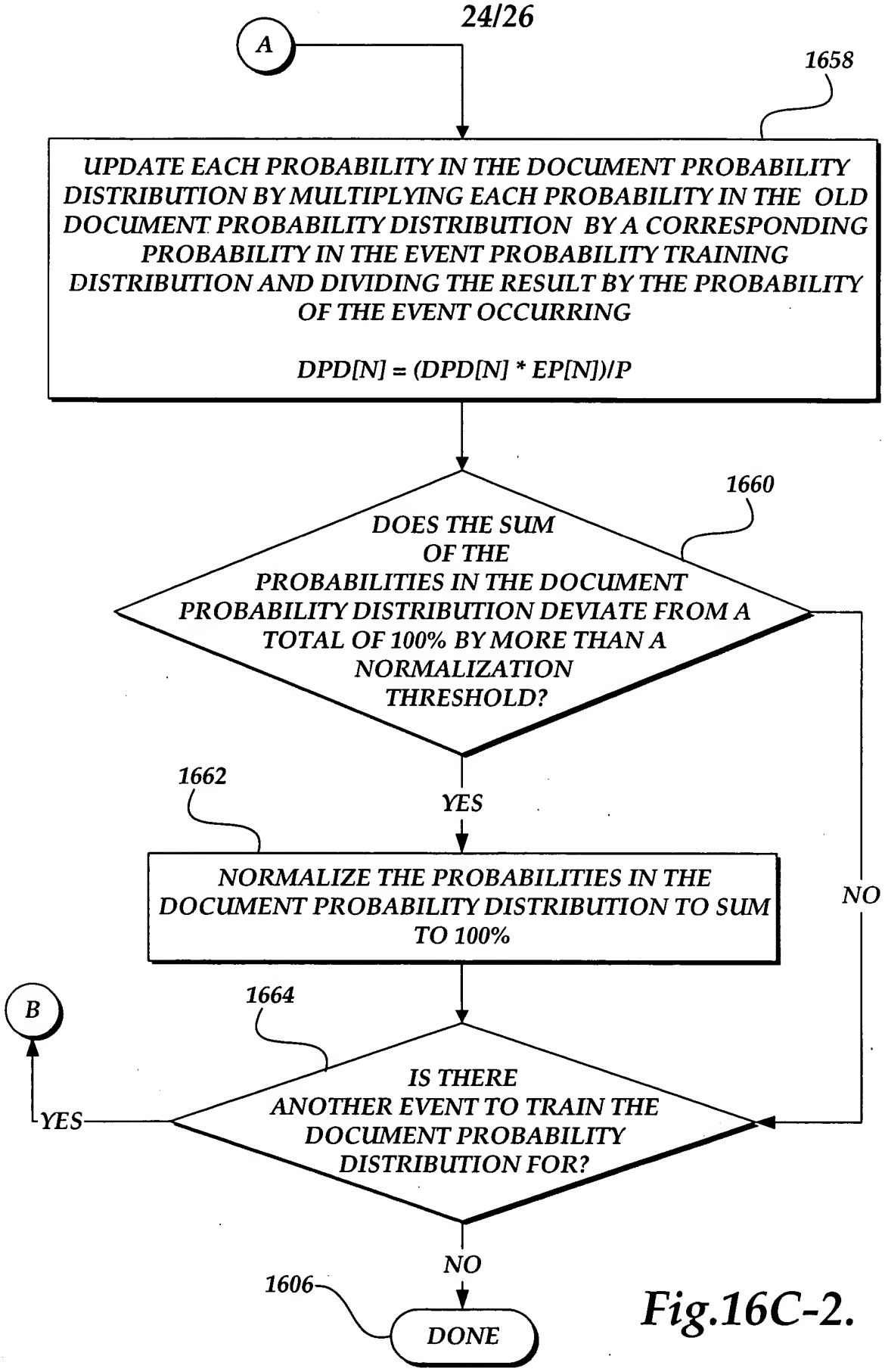


Fig.16C-2.

09487680.0274200

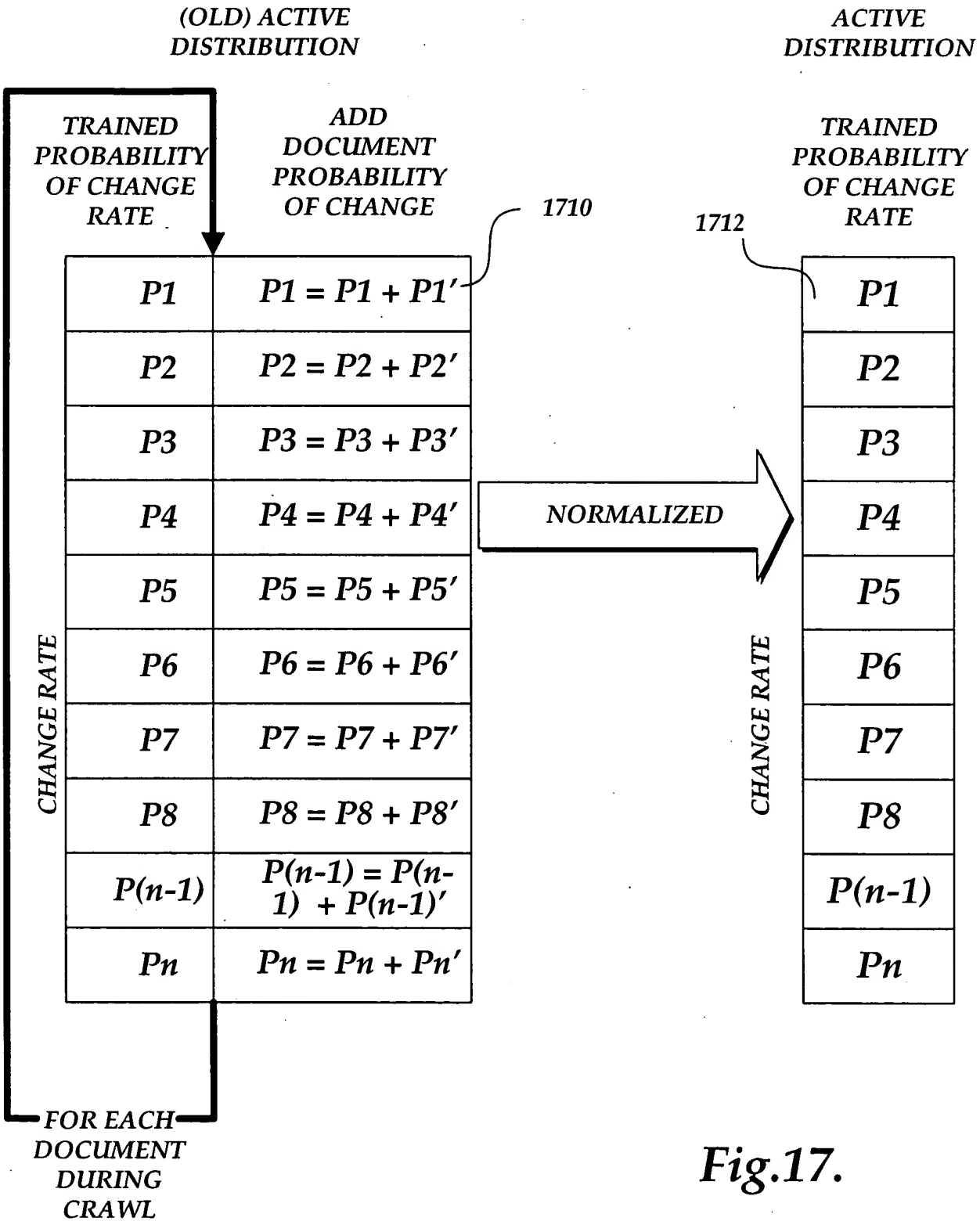

(OLD) ACTIVE DISTRIBUTION
ACTIVE DISTRIBUTION
TRAINED PROBABILITY OF CHANGE RATE
ADD DOCUMENT PROBABILITY OF CHANGE
1710
1712
TRAINED PROBABILITY OF CHANGE RATE
P1
P1 = P1 + P1′
P2
P2 = P2 + P2′
P3
P3 = P3 + P3′
P4
P4 = P4 + P4′
P5
P5 = P5 + P5′
P6
P6 = P6 + P6′
P7
P7 = P7 + P7′
P8
P8 = P8 + P8′
P(n-1)
P(n-1) = P(n-1) + P(n-1)′
Pn
Pn = Pn + Pn′
NORMALIZED
CHANGE RATE
CHANGE RATE
FOR EACH DOCUMENT DURING CRAWL

Fig.17.

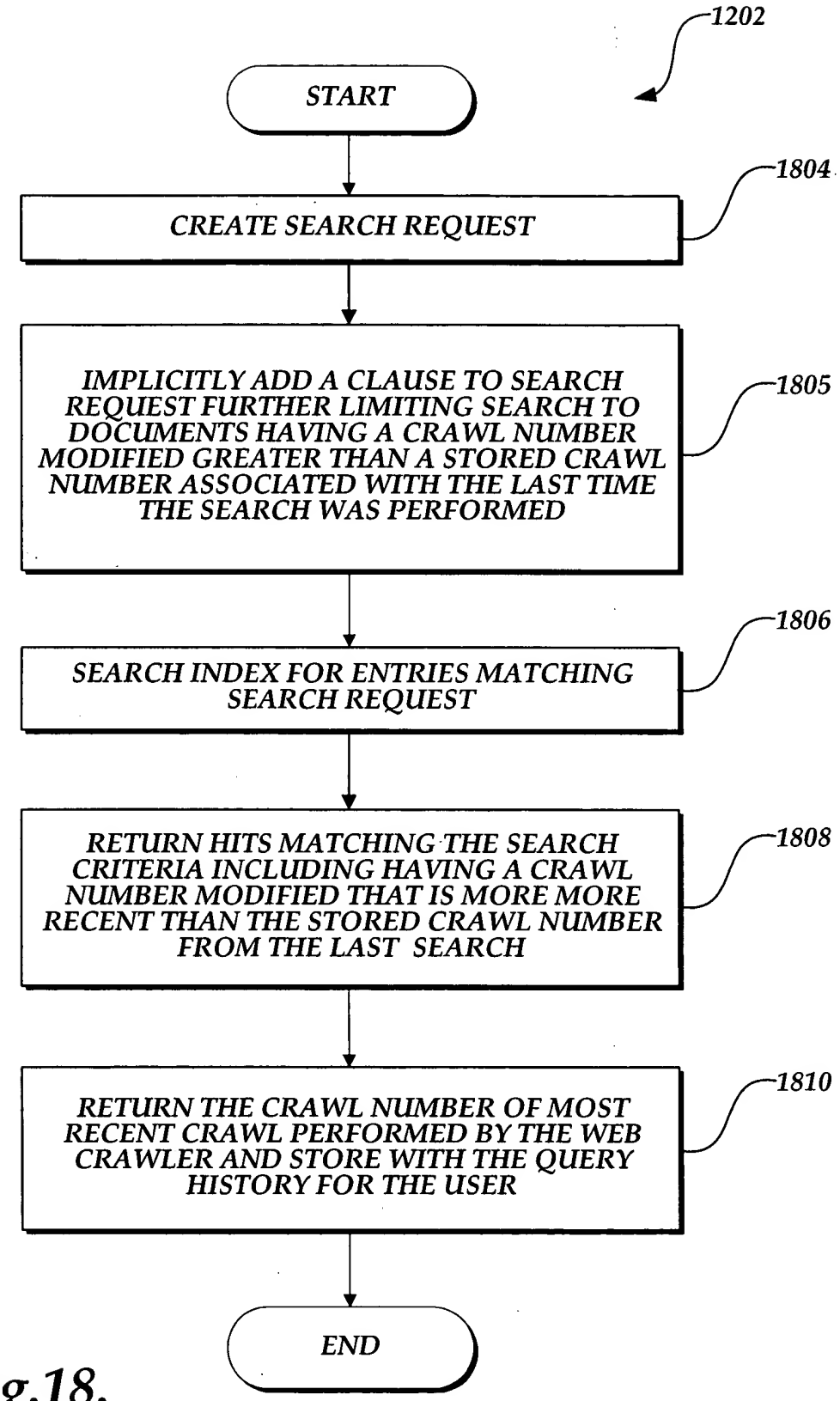


Fig.18.

0094654v1 - 8/4/2004 - 09:21:03